MULTIDISCIPLINARY SCIENCE JOURNAL

RESEARCH ARTICLE
Published Online: November 21, 2024
https://doi.org/10.31893/multiscience.2025296

Assessing decision tree, random forest, and XGBoost models for human capital readiness predictions in low-income areas

Pita Jarupunphol[a] | **Wipawan Buathong**[a] | **Suthasinee Kuptabut**[b] | **Wichidtra Sudjarid**[c]

[a]Digital Technology Program, Phuket Rajabhat University, Thailand.
[b]Computer Program, Sakon Nakhon Rajabhat University, Thailand.
[c]Environmental Science Program, Sakon Nakhon Rajabhat University, Thailand.

Abstract This article investigates applying advanced machine learning models, including Random Forest, XGBoost, and LightGBM, to predict human capital readiness in the low-income Kut Bak district. Human capital readiness, encompassing skills, knowledge, and health, is crucial for socio-economic development but remains under-researched in impoverished regions. The dataset comprises 953 individuals from 302 households, with variables including age, gender, health, employment status, and various socio-economic factors. The methodology involves rigorous data preparation, including cleaning, encoding, and normalizing, followed by feature selection and splitting the dataset into training (80%) and testing (20%) sets. The models were evaluated using a 10-fold cross-validation strategy, ensuring robust and generalizable findings. The results indicate that XGBoost outperforms the other models, achieving a mean cross-validation accuracy of 0.991, precision of 0.993, recall of 0.992, F1-score of 0.992, and an AUC-ROC score of 0.995. Random Forest follows closely with a mean accuracy of 0.987, precision of 0.990, recall of 0.988, F1-score of 0.989, and an AUC-ROC score of 0.993. LightGBM, while still performing well, shows slightly lower metrics with a mean accuracy of 0.975, precision of 0.978, recall of 0.976, F1-score of 0.977, and an AUC-ROC score of 0.980. The confusion matrix analysis reveals that XGBoost has the highest number of correct classifications, with 64 true negatives, 0 false positives, 2 false negatives, and 125 true positives. The feature importance analysis highlights 'Family Responsibilities', 'Elderly', and 'Crop Cultivation Specialist' occupation as significant predictors across models, though LightGBM emphasizes 'Age' and 'Gender'. This research underscores the utility of machine learning in socio-economic planning, offering actionable insights for policymakers aiming to enhance human capital readiness in economically disadvantaged areas. Future research should focus on expanding datasets, fine-tuning model parameters, and exploring additional socio-demographic variables for improved predictive accuracy.

Keywords: human capital readiness, machine learning, LightGBM, random forest, XGBoost

1. Introduction

The study of human capital in socio-economic contexts has garnered significant attention due to its profound implications on economic development, social equity, and individual well-being. Human capital encompasses the skills, knowledge, and health individuals accumulate throughout their lives, enabling them to realize their potential as productive members of society. While a considerable body of research has explored the determinants and impacts of human capital in affluent contexts, less attention has been directed towards low-income areas, where the challenges and dynamics can be starkly different and more complex. Limited access to education, healthcare, and economic opportunities in low-income regions can severely impede human capital development. These areas often suffer from higher unemployment rates, underemployment, and economic stagnation, further exacerbating social and economic disparities. Given these challenges, it is critical to adopt innovative and precise methods to assess and enhance human capital readiness and the preparedness of individuals to join and contribute effectively to the workforce.

Human capital readiness is a comprehensive notion encompassing a range of attributes, including skills, knowledge, and health, essential for effective workforce participation and contributing to societal and economic progress (Murfi & Hendarman, 2023). This readiness is particularly vital in low-income areas that significantly influence economic growth, social development, and efforts to reduce poverty. Challenges unique to these communities, such as restricted access to quality education, healthcare, and viable employment options, render the development of human capital both complex and critical. Research in this field has focused on devising detailed metrics to gauge human capital readiness, factoring in aspects like educational attainment, health status, skill diversity, and employment rates. One notable metric is the Human Capital Index (HCI) developed by the World Bank, which estimates the potential human capital a child born today can achieve by 18 (Human Capital Index | Data Catalog, n.d.). Empirical evidence strongly links human capital to economic enhancement (Dankyi et al., 2022; Shaban &

 Received: August 10, 2024 | Accepted: November 6, 2024

Khan, 2023; Zhang et al., 2023), with studies like that of Zhang et al. (2023) demonstrating that educational achievements and cognitive skills substantially boost economic performance. Furthermore, investments in human capital are closely tied to faster technological uptake and innovation (Al-Tit et al., 2022).

Research also illuminates the numerous challenges low-income areas face, such as poor access to essential education and healthcare services, which hinder human capital growth. Morris et al. (2018) highlighted poverty as a significant barrier to accumulating human capital. Additionally, the impact of societal and cultural elements, such as gender norms and social stratification, on human capital formation is a focal point of the study, with discrepancies in human capital accumulation often observed across different genders and social groups. The body of literature thoroughly examines various policy measures to boost human capital, including educational reforms, vocational training, and health initiatives, with scholars like Bos et al. (2024) stressing the critical importance of early childhood interventions for sustainable human capital development. Recently, the integration of big data and machine learning technologies has introduced innovative methodologies into this research area. Several machine learning models have shown promise in various applications, from predicting economic trends to optimizing resource allocation (Buchholz & Grote, 2023; Lundberg et al., 2022; Wang et al., 2024). These models can process large datasets with numerous variables, capturing complex, non-linear relationships that traditional statistical methods might miss. This current study leverages these advanced techniques to predict human capital readiness, mainly focusing on low-income regions, contributing significantly to the rich and diverse landscape of theoretical and practical knowledge on human capital readiness.

1.1. Machine Learning in Socio-Economic Research

The field of socio-economic research has witnessed a significant surge in the application of machine learning techniques, with numerous studies demonstrating the capacity of various algorithms to extract meaningful patterns from complex datasets (Avirappattu et al., 2022; Bloise & Tancioni, 2021; Kumar, 2023; Li et al., 2022b). While the literature encompasses a wide array of machine learning approaches, including decision trees (Cheng & Lin, 2024), this study focuses specifically on three advanced ensemble methods: Random Forest (Niu et al., 2020), XGBoost (Xu et al., 2024), and LightGBM (Shehadeh et al., 2021). The selection of these three models is grounded in the strengths these models offer for socio-economic predictive modeling. For example, ensemble methods like Random Forest are known for handling high-dimensional data and mitigating overfitting by averaging the predictions of multiple decision trees. This is particularly valuable when dealing with complex, multidimensional socio-economic datasets. XGBoost and LightGBM, as advanced implementations of gradient-boosting techniques, were selected for their computational efficiency and superior performance in handling structured data, which is prevalent in socio-economic studies. Both models optimize predictions by sequentially building trees and adjusting for errors from prior iterations, enhancing their accuracy in capturing the non-linear relationships in socio-economic factors. Additionally, these models incorporate built-in mechanisms for handling missing data and categorical variables, which are common in socio-economic datasets, making them highly suitable for our predictive task (Shehadeh et al., 2021; Xu et al., 2024). Other methods, such as logistic regression or support vector machines, while powerful, may not capture the complex, non-linear interactions among socio-economic variables as effectively as the chosen ensemble models. The following sections detail these algorithms, exploring their unique characteristics, applications, and performance in socio-economic predictive modeling.

1.1.1. Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training, deriving its final prediction from the most frequent output among these trees. This approach enhances predictive accuracy and addresses overfitting issues often associated with individual decision trees. The versatility of Random Forest is evident in their wide-ranging applications, from remote sensing and species distribution modeling (Hanberry, 2024) to genomics and proteomics for biological data classification. Cao et al. (2024) demonstrated their utility in identifying cancer-related microbial biomarkers, while Park et al. (2022) showcased their effectiveness in stock market trend prediction and algorithmic trading. Given its robust performance across diverse domains and its ability to handle complex, high-dimensional datasets, the Random Forest algorithm is particularly well-suited for this study's objective of predicting human capital readiness in low-income areas, where multiple socio-economic factors interact in intricate ways. The forecast is usually determined by taking the average (in the case of regression) or by selecting the majority vote (in the case of classification) from the individual trees. Equation 1 denotes the formula used for predicting a random forest.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{n} T_i(x) \quad (1)$$

Where:
$\hat{y}$ is the final prediction.
N is the number of trees in the forest.
$T_i(x)$ is the prediction of the *i*th tree for input x.

1.1.2. XGBoost

XGBoost, an advanced implementation of gradient-boosted decision trees, is renowned for its speed and performance in handling large-scale machine-learning tasks. Its versatility is evident across a broad spectrum of complex applications. Xu et al. (2024) demonstrated its efficacy in analyzing multi-modal data for self-harm behavior prediction in young adults. Abbasimehr et al. (2023) employed XGBoost to forecast solar power generation in the energy sector, highlighting its potential in optimizing renewable energy integration. The algorithm's adaptability extends to urban planning and environmental science, with applications in traffic flow optimization (Vlachogiannis et al., 2023) and air pollutant reduction (Li et al., 2022a). This broad applicability underscores XGBoost's capacity to address multifaceted challenges across diverse domains. Given its proven ability to handle complex, large-scale datasets and extract meaningful patterns from multidimensional data, XGBoost is particularly well-suited for this study on human capital readiness prediction in low-income areas. Its robust performance in socio-economic and environmental contexts suggests it can effectively capture the intricate interplay of factors influencing human capital development in resource-constrained settings. It constructs trees in sequence, with each subsequent tree aiming to rectify the inaccuracies of its predecessors. Equation 2 defines the objective function of XGBoost.

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \quad (2)$$

Where:

$l(y_i, \hat{y}_i^{(t)})$ is the loss function, measuring the difference between the actual label y^i and the prediction $\hat{y}_i^{(t)}$ at iteration *t*.

$\Omega(f_k)$ is the regularization term to prevent overfitting.

$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i)$ is the prediction at iteration *t*, summing over all the trees built so far.

1.1.3. LightGBM

LightGBM (Light Gradient Boosting Machine) is an advanced, high-performance gradient boosting framework developed by Microsoft, designed to be highly efficient, accurate, and capable of handling large-scale data with lower memory usage than other boosting algorithms (Shehadeh et al., 2021; Wang et al., 2022). It employs a novel Gradient-based One-Side Sampling (GOSS) to filter out data instances with small gradients, focusing computational resources on instances that contribute more to the information gain. This approach, combined with Exclusive Feature Bundling (EFB) for reducing feature dimensions, allows LightGBM to achieve faster training speed and higher efficiency while maintaining accuracy (Wang et al., 2022). The algorithm's tree-growing strategy differs from traditional level-wise tree growth, instead opting for a leaf-wise approach with best-first decision tree growth, which can lead to more complex trees and potentially better accuracy, especially for larger datasets. LightGBM supports various objective functions, including regression, classification, and ranking tasks, making it versatile for various machine learning applications. Its ability to handle categorical features natively without extensive preprocessing further enhances its utility in real-world scenarios where mixed data types are common (Shehadeh et al., 2021). The framework also includes built-in cross-validation and early stopping mechanisms, facilitating model tuning and preventing overfitting. These features, combined with its computational efficiency, have made LightGBM increasingly popular in academic research and industry applications, particularly in scenarios requiring rapid model iteration or large-scale datasets. Equation 3 illustrates that while the LightGBM formula resembles that of XGBoost, it places emphasis differently.

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \quad (3)$$

Where:

$l(y_i, \hat{y}_i^{(t)})$ is the loss function, similar to XGBoost.

$\Omega(f_k)$ is the regularization term, similar to XGBoost.

LightGBM uses a histogram-based algorithm for faster computation and a leaf-wise growth strategy to grow trees more efficiently.

We have explored the functions and equations of Random Forest, XGBoost, and LightGBM; Table 1 provides a detailed comparative analysis of these three machine learning models.

This research assesses the efficiency of these advanced machine learning techniques in predicting human capital readiness in the Kut Bak district in Sakon Nakhon Province, a representative low-income area in Thailand. By employing a comparative approach to assess the effectiveness of different machine learning models, this study not only identifies the most predictive factors of human capital readiness but also offers insights into the practical applications of these technologies in policy-making and development strategies. The potential impact of this research on policy-making and development strategies is significant, as it can guide the allocation of resources and interventions to improve human capital readiness in economically disadvantaged areas, thereby fostering greater economic resilience and social progress.

2. Materials and Methods

The research adopts a quantitative approach, utilizing a comparative analysis design to assess and compare the performance of Random Forest, XGBoost, and LightGBM models in predicting human capital readiness in low-income areas.

Table 1 Comparative analysis of Random Forest, XGBoost, and LightGBM: purposes and advantages.

Model	Purpose	Advantages
Random Forest	Ensemble learning for classification and regression.	Reduces overfitting compared to single decision trees. Handles large datasets with higher dimensionality. Provides feature importance rankings. Effective in remote sensing and biological data classification.
XGBoost	Optimized gradient boosting for speed and performance.	High performance and accuracy. Efficient handling of large-scale data. Built-in regularization to prevent overfitting. Handles missing data automatically Flexible for various objective functions.
LightGBM	High-performance gradient boosting framework.	Faster training speed and higher efficiency. Lower memory usage compared to other boosting algorithms. Supports categorical features natively. Leaf-wise tree growth for potentially better accuracy. Built-in cross-validation and early stopping.

2.1. Data Collection

This study utilizes three advanced decision tree-based algorithms to identify critical patterns for predicting employment status among 622 working individuals primarily engaged in agricultural activities. The dataset comprises 953 individuals from 302 households in the Kut Bak district. It includes 35 variables representing key socio-economic factors such as age, gender, health condition, employment status, family responsibilities, welfare participation, and occupational roles (including agriculture and other industries). Age and health condition are categorical variables, while gender is binary (1 for male, 2 for female). The remaining attributes are dichotomous (0 for No, 1 for Yes), covering one variable related to working status, four related to welfare support (infant care, disability assistance, government aid, and senior citizen support), and four reasons for unemployment (elderly care, disability, illness, family responsibilities). Twelve variables are related to occupational roles (e.g., crop cultivation specialist, animal husbandry expert, corporate professional, and textile creator). In comparison, eleven variables capture various skills (e.g., crop cultivation expertise, machinery operation, culinary skills, digital technology proficiency, and market analysis). The primary objective of this research is to evaluate and compare the effectiveness of the Random Forest, XGBoost, and LightGBM algorithms in predicting employment status within the specific socio-economic context of the study area.

The investigation adheres rigorously to ethical standards and has obtained formal approval from the ethics committee at Sakon Nakhon Rajabhat University. The study was granted full board review clearance (approval number HE 65-099), valid from August 31, 2022, to August 31, 2023. This ethical framework ensures that all research activities are conducted with utmost respect for the rights and well-being of the study participants. By combining advanced machine learning techniques with stringent ethical considerations, this research aims to provide valuable insights into employment patterns in the agricultural sector of the Kut Bak district, while maintaining the highest standards of research integrity and participant protection.

2.2. Data Preparation

Data compilation and aggregation from multiple sources ensured a comprehensive dataset that reflects individuals' diverse attributes and conditions in the Kut Bak district. The data underwent a rigorous cleaning process to rectify inconsistencies, remove duplicates, and handle missing values effectively. Various imputation strategies tailored to the data's nature were employed to preserve the dataset's integrity. Following the cleaning process, a transformation of the dataset aligned with analytical needs. This involved encoding categorical variables and normalizing numerical features to ensure consistency across the dataset, facilitating effective integration with machine learning algorithms. To ensure the robustness of the models, we employed a multi-step approach to feature selection and data cleaning. We initially included all available variables in the model. However, to improve the efficiency and accuracy of the machine learning models, we applied a feature importance analysis during preliminary Random Forest operations. Variables contributing little to predictive accuracy were excluded. This strategy allowed us to focus on the most relevant socio-economic predictors of human capital readiness, such as family responsibilities, elderly care, and health status.

Given the rural setting and the socio-economic context, the dataset contained several missing values, particularly in health and employment data. We addressed missing values through imputation strategies based on the mean for continuous variables (e.g., age, income) and the mode for categorical variables (e.g., gender, employment status). To ensure the reliability of the dataset, we conducted an outlier detection and removal process, eliminating extreme values that may have resulted from data entry errors or unrepresentative responses. Categorical variables were encoded using a binary scheme for the dichotomous variables and encoding for multi-class variables, such as occupation type, to facilitate their integration into

machine learning algorithms. Finally, normalization was applied to continuous variables to ensure all features were on the same scale, which is especially important for gradient-boosting algorithms like XGBoost and LightGBM. Finally, the dataset was split into training and testing sets, with 80% allocated for training to develop the models and 20% reserved for testing to evaluate model performance independently. This split was crucial for validating findings and ensuring the models could generalize beyond the training data. The dataset was also prepared strictly following ethical guidelines, ensuring the privacy and confidentiality of all participants' data throughout the research process.

2.3. Machine Learning Models

This study employed a rigorous methodological approach to ensure robust model training and evaluation. The dataset was partitioned using a train-test split strategy, allocating 80% of the data for model training and reserving 20% for final testing. This division allows for an unbiased assessment of model performance on unseen data. We chose three state-of-the-art decision tree-based ensemble algorithms for model selection: Random Forest, XGBoost, and LightGBM. Random Forest, known for reducing overfitting, constructs multiple decision trees and aggregates their predictions. XGBoost, an implementation of gradient-boosted trees, is renowned for its speed and performance, particularly in structured data problems. LightGBM, also a gradient boosting framework, is designed for efficiency and can handle large-scale data with lower memory usage. To ensure the reliability and generalizability of our models, we implemented a 10-fold cross-validation strategy during the training phase. This approach involves dividing the training data into ten equal subsets or folds. The model is then trained on nine folds and validated on the remaining fold, repeated ten times so that each fold serves as the validation set once. This methodology provides a more robust estimate of model performance by mitigating the impact of data partitioning and reducing the risk of overfitting. Using 10-fold cross-validation, combined with the train-test split, allows for a comprehensive evaluation of model stability and predictive power across different data subsets, enhancing our findings' reliability.

2.4. Model Training and Evaluation

The models will be trained using some of the collected data, with the rest reserved for validation. Cross-validation techniques will be employed to ensure a comprehensive assessment of the model's predictive capabilities and generalizability across different subsets of the data. This approach involves dividing the dataset into ten folds, training the models on a subset of folds, and validating them on the remaining fold. This process is repeated ten times, with each fold serving as the validation set once, ensuring that all data points are used for both training and validation. A confusion matrix will be utilized as an initial step in the validation process to compare actual and predicted cases, providing a precise measure of each model's accuracy. This matrix will display the number of true positives (correctly predicted 'working' instances), true negatives (correctly predicted 'non-working' instances), false positives (incorrectly predicted as 'working'), and false negatives (incorrectly predicted as 'non-working'). The accuracy percentage of each model in predicting both 'working' (1) and 'non-working' (0) cases will be compared to determine their effectiveness in classifying individuals' employment status. In evaluating classification model performance, a confusion matrix is a tabular representation of the model's predictive accuracy, as illustrated in Tables 2(a) and (b). This matrix is typically structured as a 2x2 or NxN configuration, where N corresponds to the number of classes in the classification problem. The matrix's rows denote the actual class labels, while the columns represent the predicted ones. The confusion matrix comprises several vital components that elucidate the relationship between predicted and actual class values:

1) True Positive (TP): Instances where the model correctly predicts the positive class.
2) False Positive (FP): Cases in which the model incorrectly predicts the positive class for a negative instance.
3) False Negative (FN): Scenarios where the model erroneously predicts the opposing class for a positive instance.
4) True Negative (TN): Occurrences where the model accurately predicts the negative class.

Table 2 Confusion Matrix for measuring the model's performance (a).

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Table 2 Confusion Matrix for measuring the model's performance (b).

		Predicted Class			
		C_1	C_2	...	C_N
Actual Class	C_1	$C_{1,1}$	FP	...	$C_{1,N}$
	C_2	FN	TP	...	FN
	...	...	...	...	...
	C_N	$C_{N,1}$	FP	...	$C_{N,N}$

2.5. Model Performance Evaluation

Model performance evaluation in this study encompasses a comprehensive analysis utilizing five essential metrics to assess and compare the efficacy of the Random Forest, XGBoost, and LightGBM models. These metrics include accuracy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic curve (AUC-ROC). Accuracy provides an overall measure of correct predictions, while precision and recall offer insights into the models' abilities to minimize false positives and capture all relevant cases. The F1-score, as the harmonic mean of precision and recall, provides a balanced measure particularly useful for datasets with uneven class distributions. The AUC-ROC, derived from plotting the true positive rate against the false positive rate at various thresholds, is a comprehensive indicator of each model's discriminatory capability. In this case, six equations measure the models' performance.

1. Accuracy represents the proportion of correct predictions made by the forecast model out of all predictions and can be determined using Equation 4.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (4)$$

2. Precision represents the ratio at which the model accurately identifies the specified class, divided by the total number of instances the model predicts as both true and false, as demonstrated in Equation 5.

$$Precision = \frac{TP}{(TP+FP)} \quad (5)$$

3. Recall represents the ratio of correctly predicted instances of a specific class to the total number of actual events of that class, as detailed in Equation 6.

$$Recall = \frac{TP}{(TP+FN)} \quad (6)$$

4. The F1-Score, determined by the harmonic mean of precision and recall, is used to evaluate the overall effectiveness of the model, as outlined in Equation 7.

$$F1 = 2 * \left(\frac{Precision*Recall}{Precision+Recall}\right) \quad (7)$$

5. AUC-ROC is a metric represented as a curve to display performance metrics for multiclass classification scenarios. It depicts a graph where the True Positive Rate is plotted vertically against the False Positive Rate, shown horizontally as 1 minus specificity. When the AUC value nears one, it indicates high model efficiency, demonstrating the model's ability to accurately distinguish between 'true' and 'false' outcomes.

This multi-faceted approach to model evaluation allows for a nuanced understanding of each algorithm's strengths and weaknesses. Accuracy offers a general performance overview, but precision and recall provide deeper insights into the models' abilities to handle false positives and negatives. The F1 score balances these considerations, proving especially valuable in scenarios with imbalanced classes. By summarizing the model's performance across all possible classification thresholds, the AUC-ROC provides a robust measure of each model's overall discriminative power. Collectively, these metrics enable a thorough and balanced assessment of model performance, facilitating informed decisions about model selection and potential areas for improvement in predicting employment status in the agricultural sector of the Kut Bak district.

2.6. *Model Interpretation*

This section will conduct a comprehensive comparative analysis of the Random Forest, XGBoost, and LightGBM algorithms to ascertain which most effectively identifies the determinants of human capital readiness in the Kut Bak district. This analysis will be predicated on two principal methodologies. Firstly, a feature importance analysis will be undertaken, comparing the relative significance of various predictors across the three models. This approach will elucidate the key factors driving the prediction of human capital readiness, potentially revealing insights into the socio-economic dynamics of the region. By comparing the feature importance rankings derived from each algorithm, we aim to identify consistencies and discrepancies in the models' interpretations of the data, thereby providing a nuanced understanding of the predictive landscape. Secondly, an in-depth examination of learning curves will be performed for each model. This analysis will focus on diagnosing potential issues of overfitting or underfitting and assessing each model's capacity for improvement with additional data. By scrutinizing the convergence patterns of training and validation scores across varying sample sizes, we can gauge the models' generalization capabilities and efficiency in utilizing the available data. This investigation will not only shed light on the current performance of each algorithm but also provide valuable insights into their potential for enhanced accuracy with expanded datasets.

3. Results

This section outlines the outcomes of applying machine learning models to assess human capital readiness in the Kut Bak district. It evaluates the performance of Random Forest, XGBoost, and LightGBM models, discussing their effectiveness

and implications for socio-economic development strategies. This section highlights key predictors, examines model strengths and limitations, and considers the broader impact of the findings on policy and practice.

3.1. Cross-Validation Scores

The cross-validation scores for Random Forest, XGBoost, and LightGBM models reveal distinct patterns in performance stability and overall effectiveness. Random Forest shows scores ranging from 0.980 to 0.993, with a mean of approximately 0.987, indicating a high level of accuracy with some variability in performance across different folds. XGBoost, with scores mostly clustering around 0.993 and a slightly higher mean of 0.991, demonstrates a robust performance and a higher consistency compared to Random Forest. This suggests that XGBoost might be more reliable for maintaining accuracy across different subsets of data. On the other hand, LightGBM shows the most variability in scores, ranging from 0.961 to 0.993, and has the lowest mean score of about 0.975. This indicates that while LightGBM can achieve high performance, it might be more sensitive to changes in the dataset, resulting in lower consistency across different evaluations. In summary, XGBoost appears to be the most stable and influential among the three models, with Random Forest trailing closely in accuracy but slightly more variation in scores and LightGBM showing potential for high performance but with notable fluctuations. Figure 1 compares cross-validation scores for the Random Forest, XGBoost, and LightGBM models.

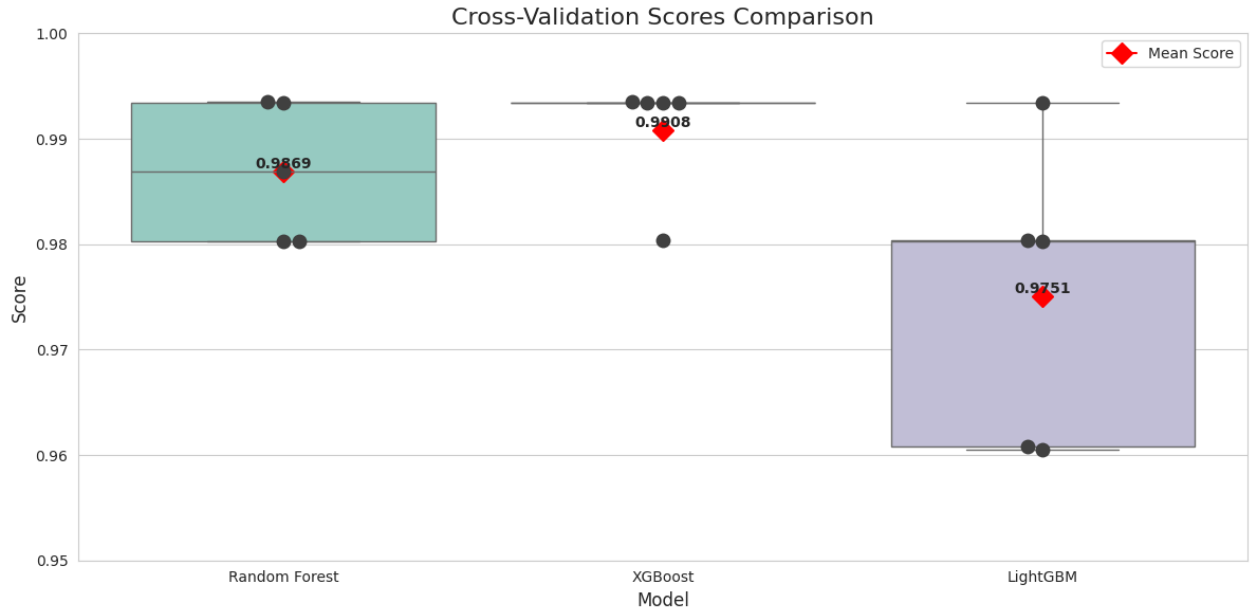

Figure 1 Cross-validation scores comparisons for Random Forest, XGBoost, and LightGBM models.

3.2. Confusion Matrix Analysis

The confusion matrix analysis for XGBoost, Random Forest, and LightGBM reveals essential insights into the models' classification performance, particularly regarding false positives (FP) and false negatives (FN). For the Random Forest model, the confusion matrix reveals 60 true negatives, 4 false positives, 2 false negatives, and 125 true positives. This indicates a high accuracy in predicting positive and negative cases, with only a few misclassifications. The XGBoost model demonstrates even stronger performance, with 64 true negatives, 0 false positives, 2 false negatives, and 125 true positives. XGBoost achieved perfect precision in identifying negative cases, though it still had two false negatives. The LightGBM model performs similarly to the Random Forest, with 60 true negatives, 4 false positives, 4 false negatives, and 123 true positives. While all models show high overall accuracy, it is crucial to interpret the socio-economic implications of these misclassifications, as they provide important indicators about the complexities of the dataset and the underlying socio-economic conditions in the Kut Bak district.

FP represents individuals classified as being 'ready' for human capital development (e.g., employed or highly skilled) when, in reality, they are not. These misclassifications may occur due to overlapping socio-economic features where individuals might possess characteristics typically associated with employment readiness, such as education or skill acquisition, but other hidden barriers (e.g., lack of job opportunities and social constraints) prevent them from participating in the workforce. The socio-economic implications of FP are significant. They suggest that while specific indicators (like education or age) may be used as proxies for readiness, they do not capture the full complexity of an individual's life situation. Misclassifying these individuals may lead to inefficient allocation of resources, such as training programs, that fail to address the natural barriers they face (e.g., caregiving responsibilities and health care access). Policymakers must ensure that interventions are targeted toward the constraints limiting workforce participation rather than relying solely on indicators that suggest readiness.

Conversely, FN represents individuals who are predicted not to be ready for employment or skills development when, in fact, they are. This misclassification could occur due to unobserved factors the model fails to capture. The socio-economic consequences of these false negatives are potentially more severe. By overlooking individuals who are genuinely ready, resources like job placement programs or skills development initiatives might bypass this group, exacerbating inequalities in access to opportunities. In particular, this FN might disproportionately affect women, youth, or individuals in non-traditional roles, as the socio-economic model may not fully capture their non-linear career trajectories or informal work contributions. These misclassifications highlight the need for models to account for more diverse data sources, such as informal sector involvement or social capital, to more accurately reflect the readiness of individuals in low-income areas. As such, both FP and FN reflect deeper issues within the socio-economic structure of the Kut Bak district. FP might indicate that traditional metrics of human capital readiness, such as education or age, are insufficient in areas where other socio-economic barriers (e.g., healthcare or family responsibilities) play a more decisive role in workforce participation. FN points to an under-recognition of informal skills or socio-cultural contributions that are not easily quantified in standard datasets but are crucial to understanding real employment readiness in low-income settings.

Comparatively, XGBoost exhibits the best overall performance with the highest number of correct classifications and the lowest number of misclassifications. It minimizes false positives, which can be crucial in many real-world applications. Both Random Forest and LightGBM show strong performance but with slightly more misclassifications than XGBoost. The Random Forest model has a slight edge over LightGBM, with fewer false negatives. These results suggest that while all three models perform well, XGBoost may be the most suitable for this classification task, especially in scenarios where minimizing false positives is a priority. However, the choice between models should also consider other factors such as computational efficiency, interpretability, and specific domain requirements. Figure 2 depicts the confusion matrix scores for the three machine learning models: a) Random Forest, b) XGBoost, and c) LightGBM, in predicting human capital readiness.

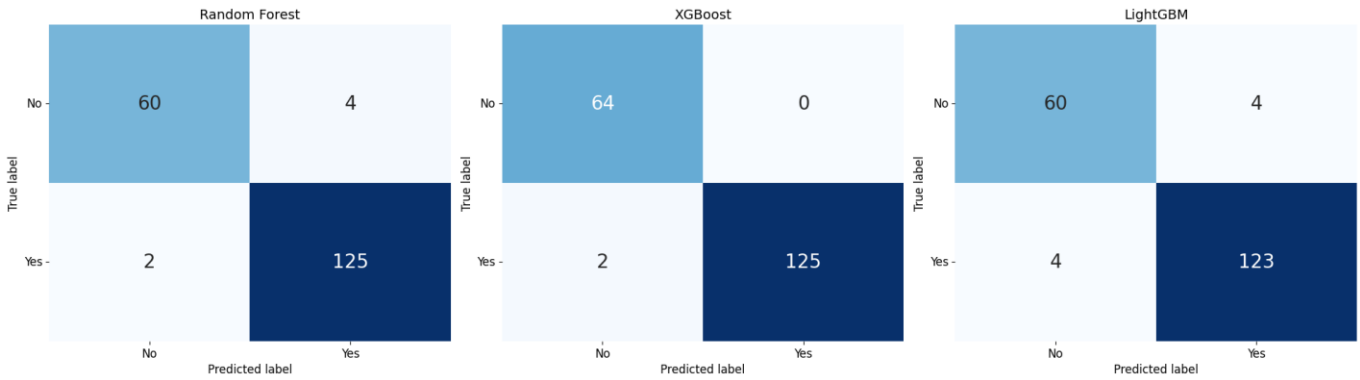

Figure 2 The confusion matrix scores for three machine learning models: Random Forest, XGBoost, and LightGBM.

3.3. Model Performance Evaluation

All three models show excellent performance, with scores consistently above 0.97 across all metrics. While XGBoost appears to be the top performer, showing the highest scores across all metrics, random forest follows closely behind XGBoost, with only slightly lower scores. Nevertheless, LightGBM, while still performing very well, shows slightly lower scores than the other two models. All models demonstrate consistent performance across metrics, indicating balanced predictions without significant trade-offs between precision and recall. In particular, the high ROC AUC scores (above 0.99) suggest that all models have excellent discriminative ability between classes. The differences in performance between the models are relatively small, particularly between XGBoost and Random Forest. Figure 3 presents a heatmap of model performance metrics.

3.4. Learning Curve Analysis

The learning curves for Random Forest, XGBoost, and LightGBM models provide valuable insights into their performance and learning patterns as training samples increase. The Random Forest model demonstrates consistently high performance across different sample sizes. With just 60 training samples, it achieves a training score of 0.998 and a cross-validation score of 0.965, indicating initial solid performance. As the sample size increases to 770, the model maintains high training scores (ranging from 0.992 to 0.995) and cross-validation scores (ranging from 0.979 to 0.986). This suggests that the Random Forest model quickly reaches a high level of performance and remains stable as more data is added, with minimal overfitting.

The XGBoost model shows a similar high-performance pattern but with a more pronounced improvement in cross-validation scores as the sample size increases. The model's performance steadily improves with a training score of 0.998 and a cross-validation score of 0.959 at 60 samples. Notably, at 770 samples, the training and cross-validation scores converge at 0.992, indicating optimal model fit without overfitting. This convergence suggests that XGBoost effectively leverages the additional data to improve its generalization capability.

LightGBM exhibits a different learning pattern compared to the other two models. It starts with lower scores at 60 samples (training score of 0.976 and cross-validation score of 0.918) but shows rapid improvement as the sample size increases. By 240 samples, its training score (0.993) surpasses that of XGBoost, though its cross-validation score (0.958) remains lower. As the sample size reaches 770, LightGBM achieves a training score of 0.992 and a cross-validation score of 0.976, indicating significant improvement in generalization. However, the persistent gap between training and cross-validation scores suggests that LightGBM might benefit from further fine-tuning or additional data to match the performance of Random Forest and XGBoost. Figure 4 displays a comparison of learning curves for the three models.

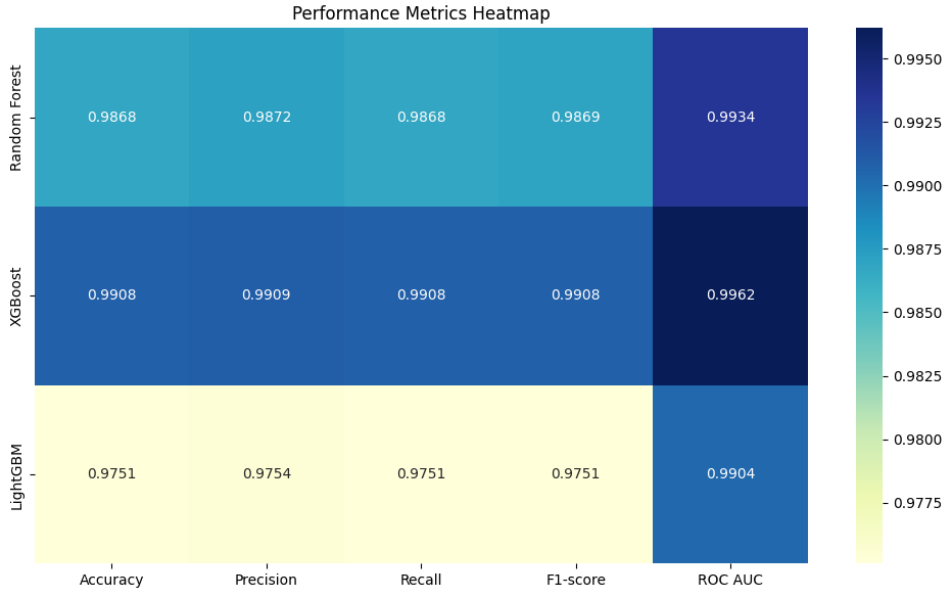

Figure 3 Model performance metrics heatmap.

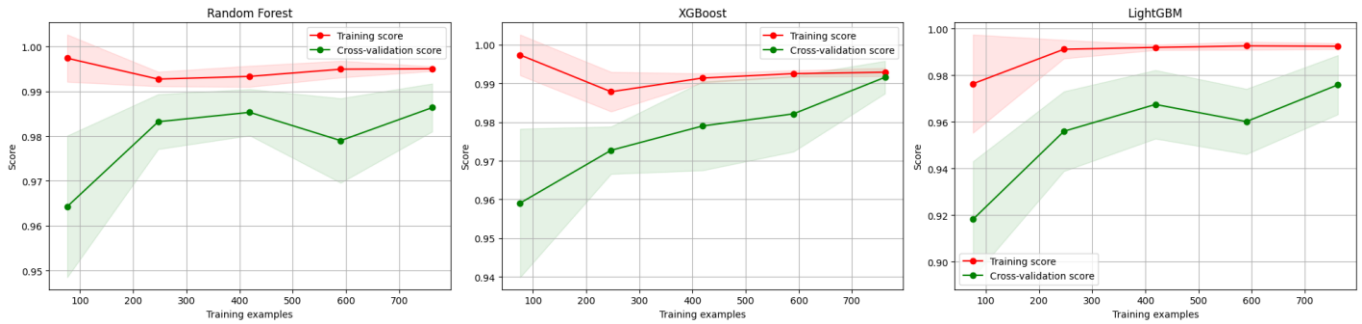

Figure 4 Learning curves of the three models.

Based on the learning curve analysis of the Random Forest, XGBoost, and LightGBM models, we can conclude that XGBoost demonstrates the best overall performance and learning characteristics for this particular classification task. XGBoost shows superior performance for several reasons. For example, XGBoost achieves perfect convergence between its training score (0.992) and cross-validation score (0.992) at 770 training samples. This indicates optimal model fit without overfitting, suggesting excellent generalization capability. In addition, XGBoost shows steady improvement in cross-validation scores as the sample size increases, from 0.959 with 60 samples to 0.992 with 770 samples. This demonstrates the effective utilization of additional data to enhance model performance. In particular, XGBoost maintains high scores throughout, with both training and cross-validation scores consistently above 0.95, even with smaller sample sizes. While Random Forest also performs admirably with high and stable scores across different sample sizes, it does not achieve the same level of convergence between training and cross-validation scores as XGBoost. LightGBM, despite showing significant improvement with increased data, still lags behind XGBoost in terms of cross-validation scores and the gap between training and cross-validation performance.

3.5. Feature Importance

The feature importance analysis across the three models reveals both consistencies and divergences in identifying key predictors for employment status. Notably, Random Forest and XGBoost demonstrate significant alignment in their top-ranked features, with ‘Family Responsibilities’ emerging as the most influential predictor in both models, followed closely by ‘Elderly’ status and ‘Crop Cultivation Specialist’ occupation. This consistency suggests a robust relationship between these factors and

employment status. However, the models differ in their weighting of age-related factors; Random Forest ranks 'Age' as the fourth most important feature, while XGBoost assigns it minimal importance. Conversely, LightGBM presents a markedly different feature importance hierarchy, with 'Age' as the most significant predictor, followed by 'Gender' and 'Health Condition' factors that receive comparatively less emphasis in the other two models.

These divergences in feature importance rankings across models underscore the complexity of predicting employment status and highlight the potential benefits of ensemble methods in capturing multifaceted relationships within the data. The prominence of family responsibilities and elderly status across Random Forest and XGBoost models suggests that socio-demographic factors play a crucial role in employment dynamics within the study area. However, LightGBM's emphasis on age and gender points to the potential importance of more fundamental demographic characteristics. The variation in feature rankings also raises important questions about model interpretability and the robustness of feature importance measures across different algorithmic approaches. Further investigation into these discrepancies could provide valuable insights into the underlying mechanisms driving employment patterns in the agricultural sector of the Kut Bak district and inform more nuanced policy interventions tailored to the specific needs of different demographic groups. Figure 5 illustrates the top 10 feature importances for the three models.

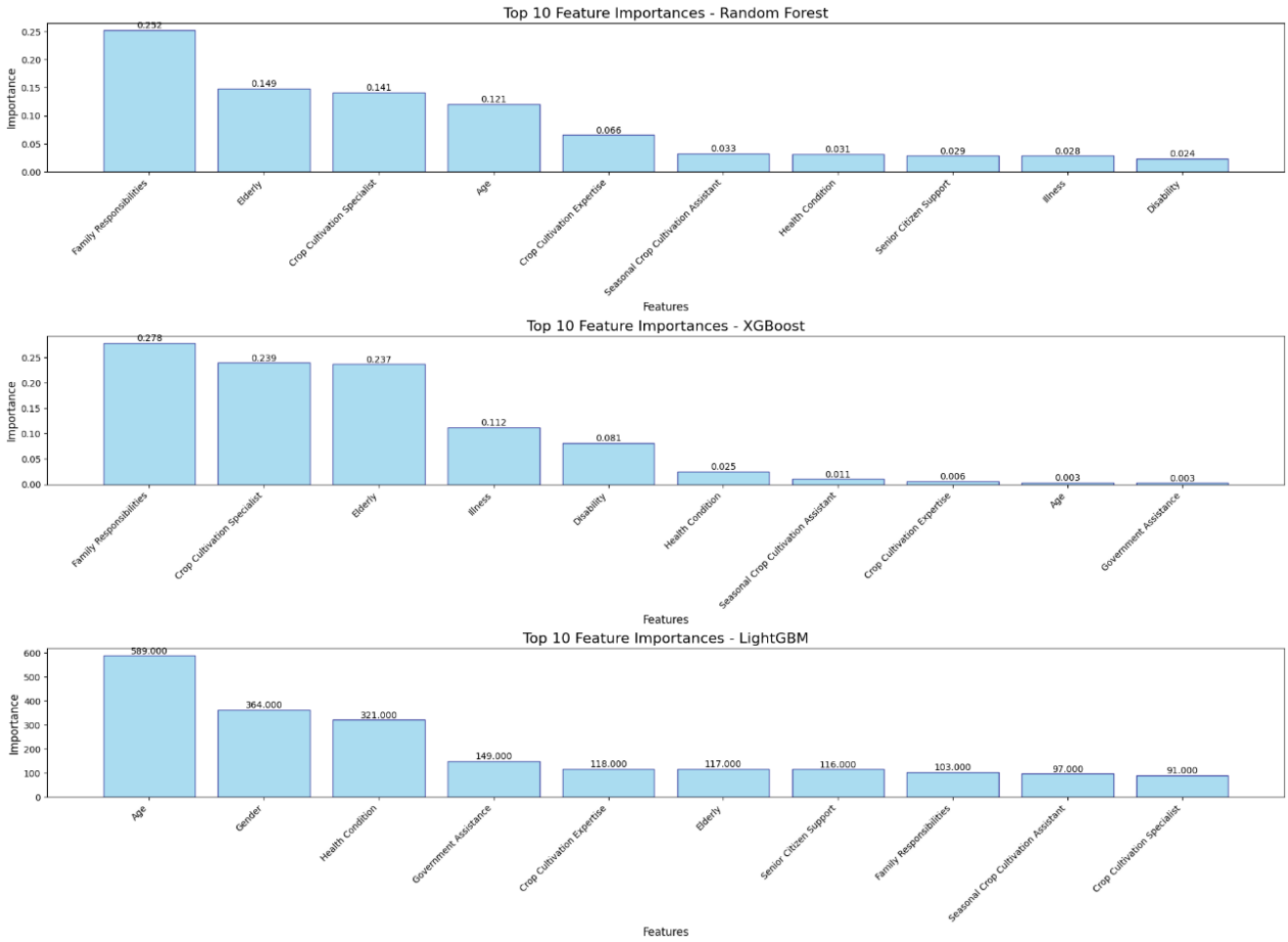

Figure 5 Top 10 feature importances of the three models.

Table 3 presents a comparative analysis of this study's three machine-learning models: Random Forest, XGBoost, and LightGBM. The table delineates each model's key performance metrics, salient features, and learning characteristics, drawing from the empirical data presented in the article. The comparative framework facilitates a nuanced understanding of the models' relative efficacies in predicting human capital readiness in the context of low-income areas. The data indicate that XGBoost demonstrates superior performance across multiple evaluation criteria, including mean cross-validation accuracy (0.991), precision (0.993), recall (0.992), F1-score (0.992), and AUC-ROC score (0.995). Random Forest exhibits comparable performance, with marginally lower metrics across all categories. LightGBM, while maintaining high-performance standards, displays more significant variability in its results than the other two models.

Notably, the confusion matrix analysis reveals that XGBoost achieves the highest number of correct classifications, with 64 true negatives 125 true positives, and no false positives. This suggests a particularly robust predictive capability for XGBoost in this specific application context. The learning curve characteristics and feature importance rankings further elucidate the

distinctive attributes of each model, providing valuable insights for model selection and optimization in similar research contexts. This comprehensive comparison not only underscores the relative strengths of each model but also highlights the importance of considering multiple performance indicators when evaluating machine learning approaches for socio-economic predictive tasks.

Table 3 Comparative analysis of Random Forest, XGBoost, and LightGBM models for human capital readiness prediction.

Aspect	Random Forest	XGBoost	LightGBM
Algorithm Type	Ensemble of decision trees	Gradient boosting	Gradient boosting
Mean Cross-Validation Accuracy	0.987	0.991	0.975
Precision	0.990	0.993	0.978
Recall	0.988	0.992	0.976
F1-score	0.989	0.992	0.977
AUC-ROC Score	0.993	0.995	0.980
Confusion Matrix Performance	60 TN, 4 FP, 2 FN, 125 TP	64 TN, 0 FP, 2 FN, 125 TP	60 TN, 4 FP, 4 FN, 123 TP
Top Feature Importance	Family Responsibilities	Family Responsibilities	Age
Learning Curve Characteristics	High initial performance, stable across sample sizes	Steady improvement, optimal convergence at larger sample sizes	Lower initial performance, rapid improvement with increased data
Performance Stability	High stability across different data subsets	Highest consistency and stability	Most variable performance
Overfitting Tendency	Low	Very low	Moderate

4. Discussion

The findings of this research show that XGBoost consistently outperforms Random Forest and LightGBM in predicting human capital readiness, achieving the highest accuracy, precision, and AUC-ROC scores. The performance of XGBoost in this study aligns with the findings of other research in socio-economic contexts that emphasize the importance of gradient-boosting algorithms in handling complex, high-dimensional data (Abbasimehr et al., 2023; Xu et al., 2024). Similar to the work of Li et al. (2022a), which highlighted XGBoost's effectiveness in optimizing urban planning and environmental predictions, our study demonstrates that XGBoost is well-suited to capturing the intricate, non-linear relationships between socio-economic variables in low-income areas. The superior performance of XGBoost in terms of both precision and recall, particularly in minimizing false positives and false negatives, supports its suitability for predictive tasks where accuracy across multiple dimensions is critical (Zhu et al., 2023). In addition, XGBoost leverages a gradient-boosting framework, which builds decision trees sequentially, with an understanding of the weaknesses in the previous ones, enabling it to capture complex relationships between variables, such as the interaction between family responsibilities, health status, and employment in predicting human capital readiness. This iterative process allows XGBoost to focus on the most challenging cases, progressively reducing bias and variance, particularly useful in complex socio-economic datasets (Xu et al., 2024). XGBoost is particularly well-suited for imbalanced datasets, which is often the case in socio-economic studies where specific outcomes (e.g., employment or educational attainment) are far less frequent.

In comparison, while Random Forest can handle class imbalance to some extent, it does not provide the same level of fine-tuning in class weight adjustments as XGBoost. The predictive accuracy of the Random Forest model in this study is consistent with findings by Hanberry (2024), highlighting its versatility and reliability in ecological and socio-economic contexts. Random Forest also performed well, showing consistent accuracy, though it lacked the fine-tuning capabilities of XGBoost in handling imbalanced classes and subtle socio-economic variations. Random Forest relies on building independent decision trees in parallel, which improves generalization and reduces overfitting but may not capture as many intricate patterns or dependencies between variables as XGBoost's sequential boosting can.

LightGBM, although also a gradient boosting framework, is optimized for speed and efficiency in larger datasets, but it might not handle overfitting as robustly as XGBoost, particularly with smaller datasets where regularization plays a crucial role. The variability in performance among the models, particularly LightGBM, which showed fluctuations as documented by Shehadeh et al. (2021), raises important considerations for their deployment. This leaf-wise growth of LightGBM allows trees to grow by optimizing loss reductions, leading to better model performance but posing an opportunity for overfitting if not correctly tuned. This variability suggests a need for further tuning and adaptation to local conditions, which may include incorporating additional socio-demographic and economic variables, as Lundberg et al. (2022) outlined in their study on machine learning applications in social sciences. Collectively, these implications direct future research towards refining machine learning applications in human capital assessments, exploring more granular data and advanced modeling techniques.

Such advancements could significantly impact policy development, aiming at sustainable socio-economic growth in low-income regions.

Furthermore, the identified socio-economic predictors, such as family responsibilities and elderly care, resonate with the broader literature on human capital development in low-income areas. For instance, Morris et al. (2018) highlighted the critical role that family obligations play in limiting workforce participation, particularly in regions with limited access to social services. Our findings are consistent with this, showing that caregiving responsibilities and age significantly influence employment readiness, as reflected in XGBoost's and Random Forest's feature importance rankings. Additionally, Bos et al. (2024) emphasize the importance of addressing formal and informal employment dynamics, stressing the need for policies that acknowledge informal work contributions. This is particularly relevant to our false negative results, where the model underpredicted employment readiness in individuals likely engaged in informal work. The literature suggests that integrating social capital and informal sector engagement variables could further enhance model performance in capturing these dynamics (Lundberg et al., 2022).

4.1. Limitations and Generalizability

While the results of this study demonstrate the efficacy of machine learning models in predicting human capital readiness, it is crucial to acknowledge the inherent limitations of applying these models in socio-economic contexts. Predicting human behavior, particularly in low-income areas, involves dealing with highly complex, non-linear relationships between variables, many of which may not be captured in available datasets. Studies have pointed out that machine learning models, including Random Forest, XGBoost, and LightGBM, often struggle with the 'black box' issue, where model predictions are accurate. Still, it lacks transparency, making it difficult for policymakers to understand the underlying decision-making process (Buchholz & Grote, 2023). This is a significant drawback when applying these models to real-world policy interventions, where interpretability is crucial.

Additionally, socio-economic data can be incomplete or biased, as the most vulnerable populations—such as informal workers or those without access to public services—are often underrepresented in datasets (Li et al., 2022b). This limitation could explain some of the false negatives in our study, where individuals who are ready for employment were misclassified as not being prepared, potentially due to the lack of data on their informal work contributions or social capital. Another critical challenge is incorporating dynamic, context-specific factors such as cultural influences, historical inequalities, or community networks into static machine-learning models (Lundberg et al., 2022). These models excel in identifying correlations but are often less effective in capturing the causal relationships and evolving socio-economic conditions that drive human behaviors. As Morris et al. (2018) argue, the socio-economic contexts of low-income areas are fluid and heavily influenced by external factors such as government policies, market fluctuations, and environmental changes, none of which are easily modeled through typical machine learning approaches.

4.2. Future Research

The study uses cross-sectional data, which does not capture the temporal dynamics of human capital readiness. In this case, longitudinal data would provide a more comprehensive understanding of the factors influencing human capital development over time. While efforts were made to address ethical concerns, the potential for biases in the dataset and the models' predictions remains a limitation that requires ongoing attention. Investigating hybrid approaches that combine the strengths of different machine-learning models could yield more robust and accurate predictions. While the models demonstrated high accuracy, several avenues for improvement could enhance their effectiveness, especially in low-income areas. For instance, future models should integrate additional variables that better capture the complexity of life in low-income areas. This could include data on informal sector participation, access to healthcare, transportation challenges, and social capital. By including these variables, models can provide a more comprehensive view of the factors influencing human capital readiness. Combining the strengths of different models, such as Random Forest and XGBoost, in a hybrid approach could improve overall predictive accuracy (Li et al., 2022). Researchers can take advantage of each algorithm's strengths while minimizing their weaknesses.

5. Conclusions

This research has successfully demonstrated the applicability and effectiveness of advanced machine learning models, including Random Forest, XGBoost, and LightGBM, in predicting human capital readiness in the Kut Bak district, a low-income area. The comparative analysis of these models has provided insights into their predictive accuracies, highlighting XGBoost as particularly effective due to its consistent performance across various metrics. The findings can provide several actionable insights and recommendations for policymakers and future research. For example, the analysis identified 'Family Responsibilities', 'Elderly Support', and 'Crop Cultivation Specialist' as key factors influencing employment status and readiness. Policymakers should prioritize interventions that address these socio-economic challenges. For example, providing childcare

support, flexible working arrangements for caregivers, and specialized training programs for agricultural workers can help improve workforce participation.

In addition, false positives in the models suggest that individuals who appear ready for employment may still face barriers such as health issues or caregiving responsibilities. Policymakers should design holistic programs offering training and employment opportunities to address these underlying socio-economic barriers. Initiatives that combine skill development with healthcare access and family support services are likely to significantly improve employment outcomes. In contrast, the false negatives in the models highlight the importance of recognizing informal skills and community contributions that are not captured by traditional readiness metrics. Policies should include programs that acknowledge and support informal workers, including certifications for informal skills and creating pathways for these individuals to enter the formal economy (Al-Tit et al., 2022).

Acknowledgment

I appreciate the *Multidisciplinary Science Journal* for providing a well-structured format for authors to follow.

Ethical considerations

I confirm that I have obtained all consent required by the applicable law to publish any personal details or images of patients, research subjects, or other individuals used. I agree to provide the *Multidisciplinary Science Journal* with copies of the consent or evidence that such consent has been obtained if requested.

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This research was supported by Program Management Unit on Area Based Development (PMU A) under contract no. A11F660115 for the project “Provinclal Poverty Alleviation Operating System Platform”.

References

Abbasimehr, H., Paki, R., & Bahrini, A. (2023). A novel XGBoost-based featurisation approach to forecast renewable energy consumption with deep learning models. *Sustainable Computing: Informatics and Systems*, *38*, 100863. https://doi.org/10.1016/j.suscom.2023.100863

Al-Tit, A. A., Al-Ayed, S., Alhammadi, A., Hunitie, M., Alsarayreh, A., & Albassam, W. (2022). The Impact of Employee Development Practices on Human Capital and Social Capital: The Mediating Contribution of Knowledge Management. *Journal of Open Innovation: Technology, Market, and Complexity*, *8*(4), 218. https://doi.org/10.3390/joitmc8040218

Avirappattu, G., Pach III, A., Locklear, C. E., & Briggs, A. Q. (2022). An optimised machine learning model for identifying socio-economic, demographic and health-related variables associated with low vaccination levels that vary across ZIP codes in California. *Preventive Medicine Reports*, *28*, 101858. https://doi.org/10.1016/j.pmedr.2022.101858

Bloise, F., & Tancioni, M. (2021). Predicting the spread of COVID-19 in Italy using machine learning: Do socio-economic factors matter? *Structural Change and Economic Dynamics*, *56*, 310–329. https://doi.org/10.1016/j.strueco.2021.01.001

Bos, J. M., Shonchoy, A. S., Ravindran, S., & Khan, A. (2024). Early childhood human capital formation at scale. Journal of Public Economics, *231*, 105046. https://doi.org/10.1016/j.jpubeco.2023.105046

Buchholz, O., & Grote, T. (2023). Predicting and explaining with machine learning models: Social science as a touchstone. *Studies in History and Philosophy of Science*, *102*, 60–69. https://doi.org/10.1016/j.shpsa.2023.10.004

Cao, L., Wei, S., Yin, Z., Chen, F., Ba, Y., Weng, Q., Zhang, J., & Zhang, H. (2024). Identifying important microbial biomarkers for the diagnosis of colon cancer using a random forest approach. *Heliyon*, *10*(2), e24713. https://doi.org/10.1016/j.heliyon.2024.e24713

Centers for Medicare & Medicaid Services (CMS), HHS (2006). Medicare program; revisions to payment policies, five-year review of work relative value units, changes to the practice expense methodology under the physician fee schedule, and other changes to payment under part B; revisions to the payment policies of ambulance services under the fee schedule for ambulance services; and ambulance inflation factor update for CY 2007. Final rule with comment period. *Federal register*, *71*(231), 69623–70251.

Cheng, C.-Y., & Lin, T.-P. (2024). Decision tree analysis of thermal comfort in the courtyard of a senior residence in hot and humid climate. *Sustainable Cities and Society*, *101*, 105165. https://doi.org/10.1016/j.scs.2023.105165

Dankyi, A. B., Abban, O. J., Yusheng, K., & Coulibaly, T. P. (2022). Human capital, foreign direct investment, and economic growth: Evidence from ECOWAS in a decomposed income level panel. *Environmental Challenges*, *9*, 100602. https://doi.org/10.1016/j.envc.2022.100602

Hanberry, B. B. (2024). Practical guide for retaining correlated climate variables and unthinned samples in species distribution modeling, using random forests. *Ecological Informatics*, 79, 102406. https://doi.org/10.1016/j.ecoinf.2023.102406

Human Capital Index | Data Catalog. (n.d.). Retrieved February 16, 2024, from https://datacatalog.worldbank.org/search/dataset/0038030/Human-Capital-Index

Kumar, S. (2023). A novel hybrid machine learning model for prediction of CO2 using socio-economic and energy attributes for climate change monitoring and mitigation policies. *Ecological Informatics*, *77*, 102253. https://doi.org/10.1016/j.ecoinf.2023.102253

Li, J., An, X., Li, Q., Wang, C., Yu, H., Zhou, X., & Geng, Y. (2022a). Application of XGBoost algorithm in the optimization of pollutant concentration. *Atmospheric Research*, *276*, 106238. https://doi.org/10.1016/j.atmosres.2022.106238

Li, Q., Yu, S., Echevin, D., & Fan, M. (2022b). Is poverty predictable with machine learning? A study of DHS data from Kyrgyzstan. *Socio-Economic Planning Sciences*, *81*, 101195. https://doi.org/10.1016/j.seps.2021.101195

Lundberg, I., Brand, J. E., & Jeon, N. (2022). Researcher reasoning meets computational capacity: Machine learning for social science. *Social Science Research*, *108*, 102807. https://doi.org/10.1016/j.ssresearch.2022.102807

Morris, M. H., Santos, S. C., & Neumeyer, X. (2018). Understanding poverty. In *Poverty and Entrepreneurship in Developed Economies* (pp. 1–20). Edward Elgar Publishing. https://doi.org/10.4337/9781788111546.00009

Murfi, M., & Hendarman, A. F. (2023). The Relationship between Human Capital Readiness in Digital Transformation Era 4.0 and Individual Performance Perception (The Case of Smart Campus in Indonesia Defense University). *American International Journal of Business Management*, *6*(3), 1–6. https://www.aijbm.com/wp-content/uploads/2023/03/A630106.pdf

Niu, T., Chen, Y., & Yuan, Y. (2020). Measuring urban poverty using multi-source data and a random forest algorithm: A case study in Guangzhou. *Sustainable Cities and Society*, *54*, 102014. https://doi.org/10.1016/j.scs.2020.102014

Park, H. J., Kim, Y., & Kim, H. Y. (2022). Stock market forecasting using a multi-task approach integrating long short-term memory and the random forest framework. *Applied Soft Computing*, *114*, 108106. https://doi.org/10.1016/j.asoc.2021.108106

Shaban, A., & Khan, S. (2023). Cultural diversity, human capital, and regional economic growth in India. *Regional Science Policy & Practice*, *15*(5), 973–992. https://doi.org/10.1111/rsp3.12528

Shehadeh, A., Alshboul, O., Al Mamlook, R. E., & Hamedat, O. (2021). Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. *Automation in Construction*, *129*, 103827. https://doi.org/10.1016/j.autcon.2021.103827

Vlachogiannis, D. M., Moura, S., & Macfarlane, J. (2023). Intersense: An XGBoost model for traffic regulator identification at intersections through crowdsourced GPS data. *Transportation Research Part C: Emerging Technologies*, *151*, 104112. https://doi.org/10.1016/j.trc.2023.104112

Wang, D., Li, L., & Zhao, D. (2022). Corporate finance risk prediction based on LightGBM. *Information Sciences*, 602, 259-268. https://doi.org/10.1016/j.ins.2022.04.058

Wang, N., Guo, Z., Shang, D., & Li, K. (2024). Carbon trading price forecasting in digitalization social change era using an explainable machine learning approach: The case of China as emerging country evidence. *Technological Forecasting and Social Change*, *200*, 123178. https://doi.org/10.1016/j.techfore.2023.123178

Xu, X.-M., Liu, Y. S., Hong, S., Liu, C., Cao, J., Chen, X.-R., Lv, Z., Cao, B., Wang, H.-G., Wang, W., Ai, M., & Kuang, L. (2024). The prediction of self-harm behaviours in young adults with multi-modal data: An XGBoost approach. *Journal of Affective Disorders Reports*, *16*, 100723. https://doi.org/10.1016/j.jadr.2024.100723

Zhang, Y., Kumar, S., Huang, X., & Yuan, Y. (2023). Human capital quality and the regional economic growth: Evidence from China. *Journal of Asian Economics*, *86*, 101593. https://doi.org/10.1016/j.asieco.2023.101593

Zhu, X., Chu, Q., Song, X., Hu, P., & Peng, L. (2023). Explainable prediction of loan default based on machine learning models. *Data Science and Management*, *6*(3), 123–133. https://doi.org/10.1016/j.dsm.2023.04.003